Profitable Prediction of Horse Racing Results using Market Price and Past Performance History

Data Analysis and Interpretation Capstone Project

3/9/2018

1 Introduction

1.1 Research Question, Explanatory Variables, and Response Variables

The purpose of this study is to determine whether the outcomes of horse races can be predicted, with a sufficient degree of accuracy for profitable betting outcomes, using variables derived from the horse's and jockey's past performance history as well as the market price as determined by the pre-race betting market. The study will be performed on a dataset of results for 28,641 race competitors or 'starters' distributed across 2,717 races at Australian racetracks in Sydney and Melbourne, gathered from 2000 through 2017.

The study aims to predict a binary response variable, **Winner**, which is 1 if the starter won the race and 0 otherwise. The explanatory variables are all quantitative, and include: (1) the final market price or dividend as determined by the betting market; (2) variables pertaining to the past performance of both the horse and jockey, such as recent finish positions, number of career starts or wins, career prizemoney, and so on. The results of the main statistical analysis (logistic regression) will be used as the basis for a wagering simulation algorithm, which can then be tested against the dataset to answer the question: if bets had been placed based on predicted probability, would the bettor have made a profit?

1.2 Rationale and Implications

The study is a collaborative project with an Australian horse-racing media company for whom I am a consultant. By applying the statistical analysis and machine learning techniques from Coursera's Data Analysis and Interpretation specialization, the current project is intended to form the basis for a new horse-racing information and prediction service which gives customers a high degree of confidence in its prospects for long-term profitability.

2 Methods

2.1 Sample

The sample includes N=28,641 starters across 2,717 races, taking place on Saturdays at metropolitan racetracks in Sydney and Melbourne, Australia, between 2000 and 2017. (I use the term 'starter' to refer to a horse competing in a particular race, as a horse may have competed in more than one race in the dataset). These criteria were chosen because Saturday metropolitan races in Sydney and Melbourne are the most popular races in Australia and have the largest betting markets. My working assumption is that larger betting markets are more accurate due to the greater number of participants, which should make it more feasible to produce a profitable predictive model.

2.2 Measures

The response variable is a binary variable, **Winner**, which is 1 if the starter won the race and 0 otherwise. The explanatory variables are all quantitative and are presented in Figure 1.

Figure 1. Explanatory Variables

Variable	Description
StartingPriceOrig	The final market price as expressed as a payout on a $1 win bet. For example, if you place a winning bet for $1 on a horse with a starting price of $2, you will receive a $2 payout
BarrierPosition	Each starter is assigned a barrier position ordered from 1 to the number of starters in the race. Lower barrier positions are preferable as they are closer to the inside of the track, and involve covering a shorter race distance.
Finish1	Recency-weighted mean of past normalized finish position.
FirstCall	Recency-weighted mean of past first call position. The first call position is the starter's position in the field at an early point in a race (depending on the distance).
CareerStarts	The horse's number of career starts prior to the race.
JockeyStarts	The jockey's number of career starts prior to the race.
JockeyWins	The jockey's number of career wins prior to the race.
WeightCarried	A 'handicapping' value assigned by the official racing administrative agency, with lower values being associated with higher winning probability.
CareerPrizeMoney	Total career prizemoney earned by the horse prior to this race.

2.3 Data Management

The SQL queries used to extract the dataset excluded races where any of the above variables could not be calculated for any of the starters in the race, so there is no further management of missing or invalid data required. For the purpose of performing Chi-Squared Tests of Independence, the explanatory variables will be binned into 4-level categorical variables by quartile. The categorical variables are identified by the suffix **_Binned**.

2.4 Analyses

Descriptive statistics are calculated for all explanatory variables, along with bar graphs showing the proportion of winners for key explanatory variables binned by quartile. Since the response variable **Winner** is binary, the bivariate analyses consist of Chi-Squared Tests of Independence performed on the binned explanatory variables, along with post-hoc analyses to determine the directionality of any statistically significant associations. The multivariate analysis is a logistic regression on the response variable **Winner**, with the result of the model being a prediction of winning probability for each starter.

The data are split into test and training sets using a 70% / 30% split performed on a *race* rather than a *starter* basis, to ensure that all starters from a particular race are in the test or training sets. The predicted winning probabilities on the test set are supplied as input to a wagering simulation algorithm to determine if the model's predictions can be used for profitable betting.

3 Exploratory Analysis

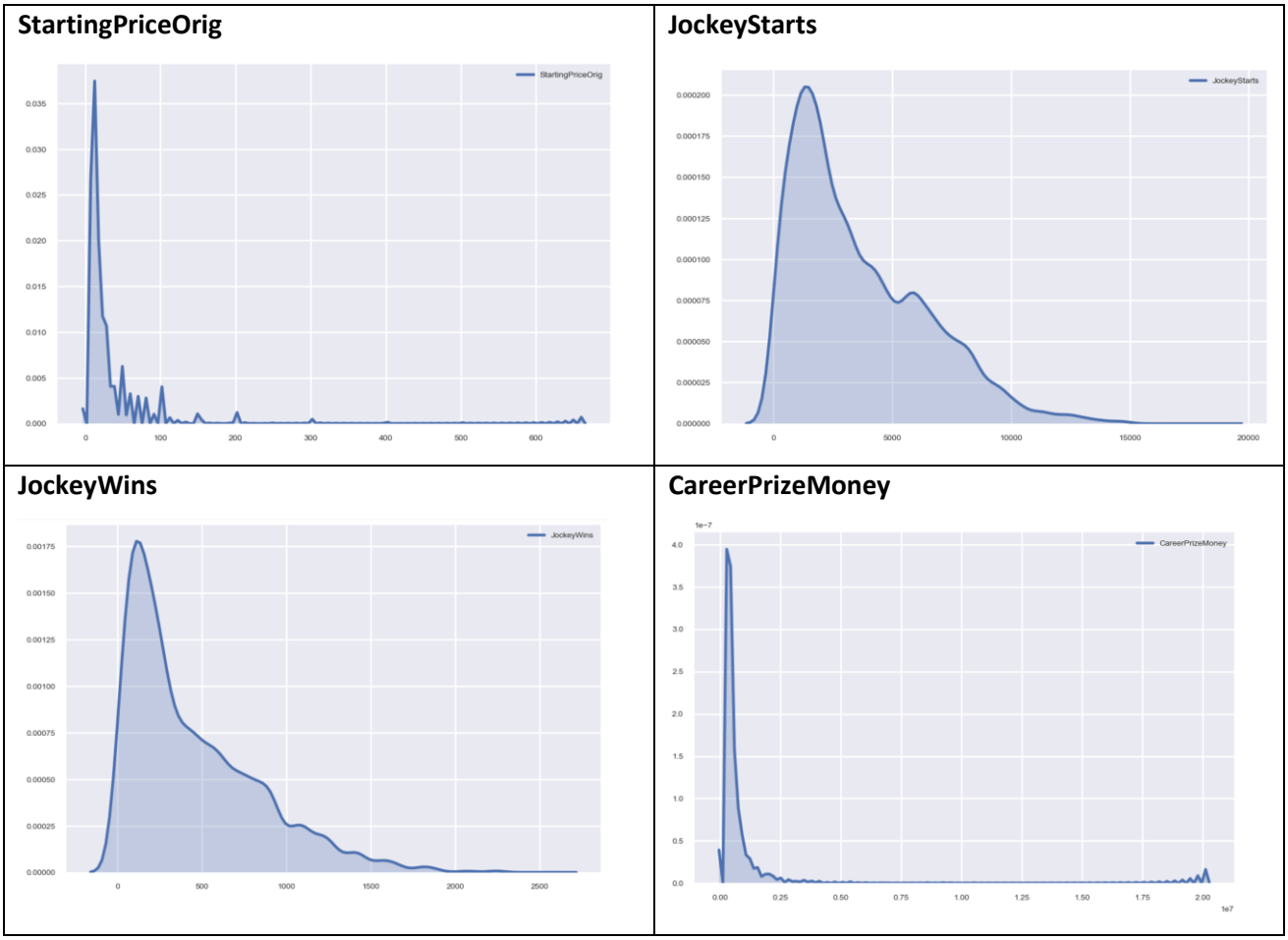
3.1 Descriptive Statistics

Figure 2 shows descriptive statistics for all explanatory variables. The mean of **StartingPriceOrig** is 21.46 (with sd = 33.16). The maximum starting price, however, is 661.00, which would indicate that the distribution of the variable is right-skewed (confirmed by the distribution plot in Figure 3). This is in accordance with common-sense bookmakers' principles – most horses are priced at low odds, with the occasional 'longshot' offered as an enticement to gamblers; odds over $100 tend to be extremely rare. Other right-skewed variables include **JockeyStarts, JockeyWins,** and **CareerPrizeMoney** (see distribution plots in Figure 3), which reflect the overall similarity of most jockey's and horse's performance histories, with a few 'champion' outliers at the tail of the distribution. These right-skewed distributions may justify the application of log transformations in regression analysis to reduce the influence of the outliers on parameter estimates.

Figure 2. Descriptive Statistics for Explanatory Variables

Variable	Count	Mean	Std	Min	25th percentile	50th percentile	75th percentile	Max
StartingPriceOrig	28,641	21.46	33.16	1.05	6.50	11.00	21.00	661.00
BarrierPosition	28,641	6.14	3.62	1.00	3.00	6.00	9.00	22.00
Finish1	28,641	0.41	0.20	0.06	0.31	0.40	0.49	5.54
CareerStarts	28,641	17.79	13.91	0	7.00	14.00	25.00	122.00
FirstCall	28,641	0.53	0.20	0.06	0.38	0.54	0.68	4.37
JockeyStarts	28,641	3,628.40	2,811.11	2.00	1,380.00	2,845.00	5,561.00	18,554.00
JockeyWins	28,641	451.24	391.05	0	139.00	325.00	680.00	2,558.00
WeightCarried	28,641	55.05	2.40	47.50	53.50	55.00	56.50	72.50
CareerPrizeMoney	28,641	200,486.85	396,732.77	0	39,285.00	98,370.00	213,650.00	20,232,950.00

Figure 3. Distribution plots for right-skewed explanatory variables



3.2 Bivariate Analyses

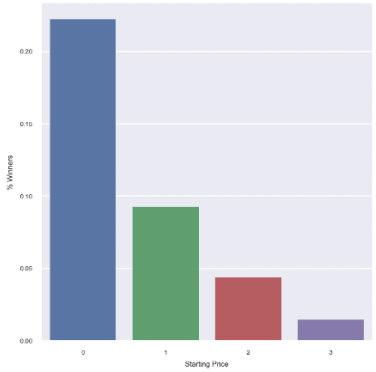
Each explanatory variable was binned into 4-level categorical variables based on quartile. Chi-Squared Tests of Independence were performed on the binned variables with **Winner** as the response variable (Figure 4). All of the Chi-Squared tests were statistically significant at the $p < 0.05$ level, with the binned variables **StartingPriceOrig_Binned** ($p = 0$), **Finish1_Binned** ($p = 4.52E\text{-}125$) showing the strongest correlations with winning probability. Bar charts for these variables (Figure 5) suggest that both **StartingPriceOrig** and **Finish1** are linearly, negatively correlated with winning probability; pairwise post-hoc comparisons for these two variables were all significant at the Bonferroni-adjusted level of $p < 0.0083$.

Figure 4. Chi-Squared Tests of Independence

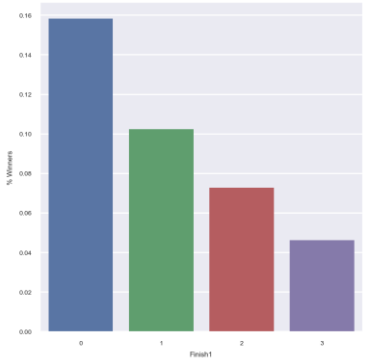
Variable	Chi-Squared Value	Chi-Squared P Value
StartingPriceOrig	2,142.59	0.00E+00
BarrierPosition	30.87	9.04E-07
Finish1	578.54	4.52E-125
CareerStarts	137.48	1.32E-29
FirstCall	111.85	4.39E-24
JockeyStarts	7.88	4.85E-02
JockeyWins	21.58	7.96E-05
WeightCarried	71.22	2.34E-15
CareerPrizeMoney	39.71	1.23E-08

Figure 5. Bar Charts Showing Percentage of Winners by Quartile

StartingPriceOrig



Finish1



4 Multivariate Analysis

4.1 Logistic Regression Model

After splitting the data into training and test sets using a 70% / 30% random split, a logistic regression model was fitted to the training set (N=19,626 starters across 1,867 races). The results of the initial logistic regression analysis are shown in figure 6. The variables **StartingPriceOrig**, **Finish1**, **FirstCall**, **JockeyStarts**, **JockeyWins**, and **CareerPrizeMoney** are all statistically significant at the $p < 0.05$ level, while **BarrierPosition**, **CareerStarts**, and **WeightCarried** are no longer statistically significant after controlling for the other variables in the model.

The next step was to make iterative refinements to the model by removing statistically insignificant variables and performing other transformations suggested by the bivariate analyses in section 3. The biggest impact on model accuracy comes from performing a log transformation on **StartingPriceOrig**, which raises the pseudo-R^2 value from 0.1172 to 0.1460 and log-likelihood from -5,444 to -5,267. Log transformations were attempted on other right-skewed

variables such as **JockeyStarts** and **CareerPrizeMoney** but did not improve model accuracy.

The last step in the refinement of the model was to remove statistically insignificant variables. The final model is given in figure 7. The complexity of the model is substantially reduced – it contains only the log-transformed **StartingPrice**, **FirstCall**, **JockeyStarts**, and **JockeyWins** – with only minimal impact on overall accuracy metrics as compared with the intermediate model, with pseudo-R^2 = 0.1453 and log-likelihood of -5,272.

Figure 6. Initial Logistic Regression Results

Logit Regression Results

Dep. Variable:	Winner	No. Observations:	19626
Model:	Logit	Df Residuals:	19616
Method:	MLE	Df Model:	9
Date:	Thu, 08 Mar 2018	Pseudo R-squ.:	0.1172
Time:	21:34:51	Log-Likelihood:	-5444.4
converged:	True	LL-Null:	-6167.4
		LLR p-value:	8.578e-306

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-0.4562	0.638	-0.715	0.475	-1.706	0.794
StartingPriceOrig	-0.1106	0.005	-21.551	0.000	-0.121	-0.101
BarrierPosition	0.0004	0.007	0.053	0.957	-0.014	0.015
Finish1	-0.6702	0.197	-3.408	0.001	-1.056	-0.285
CareerStarts	-0.0019	0.002	-0.844	0.399	-0.006	0.003
FirstCall	-0.4029	0.132	-3.047	0.002	-0.662	-0.144
JockeyStarts	-0.0001	3.12e-05	-3.775	0.000	-0.000	-5.66e-05
JockeyWins	0.0008	0.000	3.567	0.000	0.000	0.001
WeightCarried	-0.0007	0.012	-0.056	0.955	-0.023	0.022
CareerPrizeMoney	1.652e-07	5.01e-08	3.297	0.001	6.7e-08	2.63e-07

Figure 7. Final Logistic Regression Results After Model Refinement

Logit Regression Results

Dep. Variable:	Winner	No. Observations:	19626
Model:	Logit	Df Residuals:	19621
Method:	MLE	Df Model:	4
Date:	Thu, 08 Mar 2018	Pseudo R-squ.:	0.1453
Time:	21:37:24	Log-Likelihood:	-5271.6
converged:	True	LL-Null:	-6167.4
		LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	0.9834	0.103	9.530	0.000	0.781	1.186
numpy.log(StartingPriceOrig)	-1.3859	0.040	-34.780	0.000	-1.464	-1.308
FirstCall	-0.4023	0.128	-3.140	0.002	-0.653	-0.151
JockeyStarts	-7.179e-05	3.17e-05	-2.267	0.023	-0.000	-9.72e-06
JockeyWins	0.0005	0.000	2.096	0.036	2.98e-05	0.001

4.2 Wagering Simulation

The final model, fitted against the training set, was used to generate predictions of winning probability for the N=9,015 starters across 850 races in the test set. These predictions were then used as input for a wagering algorithm. The generated probabilities required adjustment in order to be useable, as the logistic regression model is fitted to the entire dataset rather than to individual races. (This limitation, and other statistical models which could be used to overcome it, is discussed in section 5). The total probabilities for each race were summed, and the adjusted probability for each starter in a particular race was defined as the original model probability over the sum of probabilities for the race. This has the desired result that the probabilities for all starters in each race sum to 1.

The wagering algorithm starts with an initial 'bank' of 1,000 units. It loops through all 850 races in the test set and places a 'bet' for any starter where the expected return is greater than 1, where expected return is defined as:

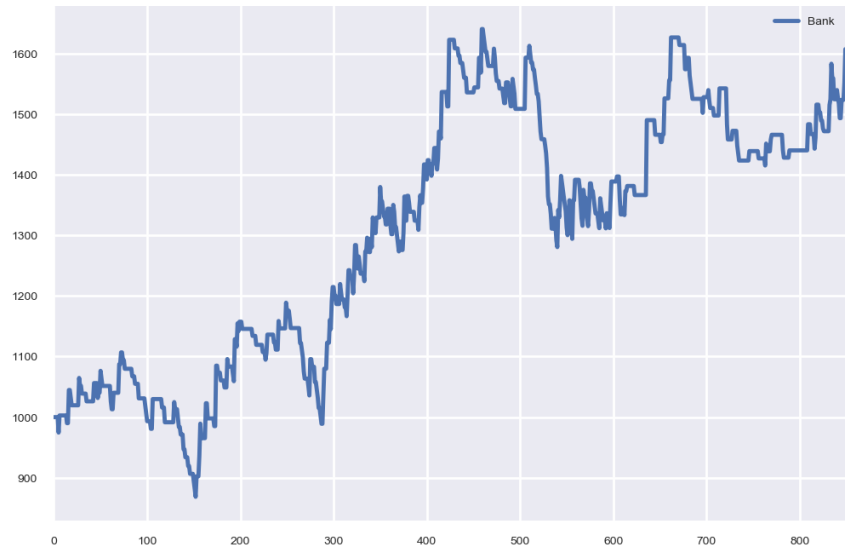
Expected Return = Starting Price x Adjusted Probability

Bet size is defined as a base bet size of 10 units multiplied by the exponential function applied to adjusted probability – in other words, bet size is increased exponentially as the model's confidence in the starter's chances increases:

Bet Size = Exp(Adjusted Probability) x Base Bet Size

The results of the wagering simulation on the test set were very promising: the wagering bank increased from 1,000 units to 1,607. Across the 850 races, 513 bets were placed for 121 winners, with an average payout of $5.07 and a winning percentage of 23.59%. Figure 8 plots the progress of the bettor's bank through the wagering simulation.

Figure 8. Progress of bettor's bank in wagering simulation



5 Conclusions and Limitations

5.1 Key Findings and Implications

This study provides evidence for the hypothesis that the prediction of horse-racing results is possible at a sufficient level of accuracy for profitable betting. The most powerful predictor of winning outcomes is the starting price, but the addition of a few statistically significant past performance factors can help in the development of a profitable predictive model. It is worth noting that the estimation of winning probability of the model does not need to be perfect – in fact, the win percentage in the wagering simulation of 23.59% is actually lower than what would be achieved by simply backing the favorite in each race (approximately 30%). The key to profitability is the fact that the probability of the winners chosen by the model was underestimated by the other bettors in the market, resulting in higher payouts.

5.2 Limitations and Future Directions

The statistical model used in this study – simple binary logistic regression – has the limitation that the likelihood estimation for each starter is calculated against the entire dataset. In effect, we are estimating the probability of each starter winning *versus all other horses in the dataset*. This assumption is not accurate – each horse only requires greater ability *than the other horses in the race* in order to win. For likelihood estimation amongst competitors within a race, we would need to fit what is called a *conditional logit* model (see Bolton and Chapman [1986] and Benter [1994]). There is no such model currently available in Python's standard statistical libraries **statsmodels** or **scipy**. A Python library for fitting conditional logit models, **PyLogit**, was very recently developed in 2016 by Timothy Braithwaite (https://github.com/timothyb0912/pylogit). Fitting a conditional logit model to the dataset using **PyLogit** and comparing the results with the simple binary logistic regression model would be an interesting direction for future research.

Bibliography

Bolton, Ruth N. and Randall G. Chapman (1986). "Searching For Positive Returns at the Track: A Multinomial Logit Model for Handicapping Horse Races." Management Science, 32, 8 (August), 1040-60.

Benter, William (1994). "Computer-Based Horse Race Handicapping and Wagering Systems: A Report." In Nick Mordin, Winning Without Thinking, Aeschelus Press, London, 2003.